DOCUMENT RESUME

ED 386 025 FL 023 161

AUTHOR Russikoff, Karen A.
TITLE A Comparison of Writing Criteria: Any Differences?
PUB DATE Mar 95
NOTE 9p.; Paper presented at Annual Meeting of the Teachers of English to Speakers of Other Languages (29th, Long Beach, CA, March 28-April 1, 1995). For related work, see ED 370 376.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *English (Second Language); Graduate Students; Higher Education; *Holistic Evaluation; Scoring; Second Language Instruction; Second Language Learning; Statistical Analysis; *Student Evaluation; *Writing Instruction
IDENTIFIERS *California Polytechnic State University

ABSTRACT

This paper offers an examination of the holistic assessment used for the Graduation Writing Test (GWT) at California State Polytechnic University at Pomona for nonnative speakers of English. Holistic assessment is a widely accepted method of evaluating student writing at the university level for administration, placement, proficiency, and certification of competency, although the validity of interpretation and application of scoring criteria have been unconfirmed for nonnative-speaking students. Similarities and differences between coursework and the writing test were examined. The constructs for comparison were operationalized for the curriculum through the criteria that faculty used for assessing English-as-a-Second-Language (ESL) writing in academic coursework, gathered by a faculty survey, and for the GWT, using holistic criteria that raters use for the high-stakes evaluation that permits, delays, or denies university graduation. Using the GWT cut-off score of 6 and below as a failing score and 7 and above as a passing score on a cumulative 12-point scale with two raters, approximately the same proportions for pass/fail were maintained. The revised scoring version, or analytic rubric, is appended. (Contains 32 references.) (NAV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

A Comparison of Writing Criteria: Any Differences?

A Paper Presented at TESOL '95, Long Beach, CA
by Karen A. Russikoff, Ph.D., English & Foreign Languages Dept.
California State Polytechnic University, Pomona

ED 386 025

Since the early 1970s, holistic assessment has become a widely accepted method of evaluating student writing at the university level for a variety of purposes, including admission, placement, proficiency and certification of competency for graduation. At present, it is one of the methods used by eighteen of the twenty campuses of the California State University system for evaluating student essay tests used to satisfy the mandated Graduation Writing Assessment Requirement (GWAR) (Borowiec, 1988; *CSU Survey*, 1990). The underlying implicit assumption for advocating the use of this assessment tool is that, reciprocally, the curriculum will prepare students for the writing test, and the backwash effect of the faculty-involved scoring of the writing test will set standards to improve the overall quality of student writing in academic coursework.

By contextualizing such testing within its educational setting, it is possible to assess the validity of the holistic assessment method used to evaluate student writing. Indeed, the validity of such scoring, specifically the interpretation and application of scoring criteria, had been previously unconfirmed for nonnative-speaking students in general (Henning, 1991), and for ESL students at Cal Poly Pomona.

The examination of the holistic assessment used for the Graduation Writing Test at Cal Poly Pomona has been necessitated by the growing population of California State University students for whom English is not a native language and who face discrimination for two immediate reasons related directly to their language backgrounds when they sit for the Graduation Writing Test:

1.) Criteria and Application: Considerable variability may exist in the manner by which holistic assessment criteria are applied by faculty raters scoring the Graduation Writing Test (GWT). These criteria may be overly broad and, as such, dependent upon a number of inconsistent and intangible variables, including rater background, variable effects of holistic training, and individual interpretation of the rubric. In spite of claims made for this form of evaluation–that it assesses the whole and not the parts–the use of holistic criteria may emphasize one criterion over another, thereby unduly penalizing ESL student essays when they are included in a mixed scoring with native-speaking student writing, as they are in GWT testing at Cal Poly Pomona; and

2.) Time and Support: A serious disadvantage in time and opportunity may exist for ESL student writers. Prior to university admission, native speakers of English generally have twelve or more years to learn and practice English academic writing, followed by validation in university coursework; conversely, ESL students have considerably less time with limited direct attention to their academic literacy needs (Cumming, 1990). Research indicates that seven to nine years is the required length of time for nonnative-speaking students to attain cognitive academic language proficiency, the level required for university coursework (Cummins, 1979; Collier, 1994) and that, for ESL students, the academic demands of coursework are the major, often only, means of developing English writing skills (Bereiter, 1980; Mendelsohn & Cumming, 1987; Hamp-Lyons, 1991; Leki, 1992). Consequently, the relationship between University curriculum and GWT testing may be all the more profound for ESL students who, by virtue of their limited time at the University, may be disproportionately dependent on their regular coursework for academic discourse opportunities as a means of GWT preparation.

Thus, to confirm the valid use of holistic assessment of nonnative-speaking students by the Graduation Writing Test, a clear understanding of the similarities and differences represented by coursework and by the writing test was necessary. The constructs for comparison were operationalized for the curriculum through the criteria which faculty use for assessing ESL writing in academic coursework (gathered by a survey of the entire faculty), and for the GWT, by the holistic criteria which raters use for the high-stakes evaluation that permits, delays or denies University graduation.

FL 023 161

BEST COPY AVAILABLE

© K. A. Russikoff, 1995

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Karen A. Russikoff
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Section I: Creation of the Analytic Rubric

Through analysis of questionnaire responses to criteria used by faculty in content courses, a new scoring guide, an analytic rubric, was created. This type of scoring guide provides distinct advantages over the general holistic scoring guide used presently for GWT assessment: that is, students readily recognize specific problems within their writing which cause them to be penalized; raters understand the specific guidelines and proportional weights for each criterion factor they are applying; and no single factor exerts disproportionate influence over passing or failing scores. (Research has noted that for nonnative student writers taking a test under timed and tense conditions, minor grammatical errors could be sole reason for the failing results [Fein,1980; McGirt, 1984; Ross et al., 1984], thereby exerting a disproportionate influence over holistic scoring.)

From the Likert-type scale requesting faculty frequency of commenting on or correcting ESL student writing, twenty criteria were mathematically weighted and derived from column score totals. Overall percentages for each of five factors were rounded off to determine the resulting rubric: Content 25%; Organization 20%; Vocabulary 15%; Language Use 20%; and Mechanics 20%. Each factor was designated levels to assist readers in scoring: High, middle and low levels were labeled, "Excellent to Very Good," "Good to Average," and "Fair to Poor," and the middle level within each factor was designed to divide numerically at the midpoint in its range.

Using the Graduation Writing Test cut-off score of 6 and below as a failing score and 7 and above as a passing score on a cumulative 12-point scale with two raters, the cut-off for the analytic scale was set at the same midpoint for pass/fail: The combined 200-point scale was set at failing with 50% (i.e., 100 points) and below and passing at over 50% (i.e., 100.5 and above). Thus, approximately the same proportions for pass/fail were maintained.

Section II: Re-scoring of the Graduation Writing Test

The Essays

The re-scoring was conducted in order to compare faculty readers' application of the two sets of criteria, those made explicit by faculty through their questionnaire responses to criteria used to evaluate coursework writing of ESL students and compiled as analytic criteria, and those used on the GWT which are broad holistic criteria. The comparisons consequently provided data to calculate differences in pass/fail ratios for ESL student essays by the application of the analytic and the holistic assessment instruments.

For the purposes of this research, the campus Testing Office provided anonymous GWT essays which were previously scored. Because no records are maintained as to test-takers' language background in connection with their essays, it was not possible to identify which essays were written by nonnative speakers. Therefore, a panel of three ESL experts judged which were authored by ESL students in order to obtain the eventual corpus of 100 student essays, covering the entire range of scoring possibilities. In addition, because few essays written by ESL students in the upper range (i.e., scores of 8, 9, 10) were made available by the Testing Office or were identified as ESL-authored by the ESL panel, essays judged to have been written by native speakers were used to complete this end of the distribution (i.e., eleven NS essays with scores of 8, 9 and 10). The final corpus of 100 previously-scored GWT essays included 89 essays judged to have been written by ESL students and 11 essays judged to have been written by native-speaking (NS) students.

The request for essays by specific scores emphasized the midrange, considered most problematic for nonnative-speaking student writers for several reasons:

1.) Essays which fall in the midrange are often most difficult for readers to assess since they usually contain characteristics of high and low levels of writing (Hamp-Lyons, 1991; Elbow, 1993);

2.) The GWT rubric's broad descriptor states that the midscore (3) applies to "papers...marred by more than a few minor grammatical inconsistencies," thus, creating a band for scoring assignment that is wider than other bands, resulting in uneven calibration of the assessment instrument (Davidson,1991), thereby potentially drawing a disproportionate number of ESL essays; and

3.) Even though the 3 score appears to be a bottom-half score, it is actually the midpoint due to the hyphenated 4-5 single score; consequently, the assignment by two raters of the 3 score, even though a midpoint in the overall range, constitutes a failing score.

© K. A. Russikoff, 1995

All test papers were from a single test administration and were written in response to the same topic prompt, one which generated personal expository prose. The selected student essays were coded and randomly assigned to one of four groups, W, X, Y, and Z. The codes were such that the readers were unaware of the original score. Four batches of 25 essays each included a range of essays meant to approximate a normal distribution with the greatest emphasis on the 6 score along with the two adjacent total scores, 5 and 7.

Readers

Since research indicates that readers' backgrounds, including professional and holistic scoring experience may affect readers' use of criteria for scoring students essays (e.g., Brown, 1991), two different groups of four readers from the University participated in re-scoring the previously-scored GWT essays (total, 8 readers). The first group was purposely selected to represent readers who regularly score the GWT but have no special expertise with ESL student writing, and the second group was purposely selected to represent readers who regularly grade ESL writing. Because the GWT Raters were faculty who regularly or frequently score the GWT, they were not asked to re-score holistically since it was assumed that their reading would only confirm prior scoring. These readers scored using only the newly devised analytic rubric. With research indicating that ESL specialists may be more sensitive to ESL writing problems, a second group of readers was included to see if this reading would be comparable to the GWT Raters. ESL Specialist readers were asked to read and score duplicate essays with the two different evaluation instruments, holistic and analytic scoring guides. Reliability of readings ranged from .65 to .875, without any norming sessions. Once they were normed to the new scoring guide, all readers agreed that their reading took no longer than holistic reading.

Analyses

Quantitative analyses were used to compare holistic scoring and analytic scoring for both GWT Raters and ESL Specialists. Correlations were used to compare overall scoring with both groups of raters and to determine interrater reliability coefficients. Simple linear, multiple, and stepwise regression analyses were used to compare holistic and analytic criteria with both groups of raters and to determine which analytic criteria most influenced each type of scoring. In addition, paired t-tests analyzed the statistical significance between groups and scoring measures. Chi square was used to assess the difference in pass/fail ratios between groups, and descriptive statistics were used to compare the applications of analytic scoring factors and differences in pass/fail ratios for each rater group. (Probability level for all analyses was set at $p<.05$.)

A lack of faculty consensus is reflected in the differences between criteria used to assess ESL student writing in regular academic courses and the criteria used to assess ESL student writing on the GWT. In course writing, faculty expect Content to be the most important indicator of competence, followed by Organization, Language Use, and Mechanics at approximately the same degree of emphasis, with Vocabulary as the least important. On the other hand, the GWT essays, scored as final products, attracted readers to Language Use (the only factor which reliably predicted holistic scores in all analyses across both analytic and holistic instruments and with both sets of readers); therefore, the "grammatical inconsistencies" of the midpoint 3 score, appear to attract a disproportionate amount of attention as faculty readers attribute more to the factor which focuses on grammar than to the others. The Graduation Writing Test appears to be testing grammar (i.e., Language Use) over all other possible factors. Consequently, as a test of grammar, this focus has a predictable penalizing effect on ESL student writers.

In addition, ESL students are affected in a practical sense by the penalizing effect of the difference in the application of scoring criteria. Considerably more essays were passed using the analytic scoring guide than were passed using the GWT guide. Since 32 additional ESL essays were passed with the paired results of analytic scoring by the GWT Raters and 51 additional essays were passed with the paired results of analytic scoring by ESL Specialists (each from the original corpus of 100 essays with an original ratio of 72 fail:28 pass), the difference between course criteria, used in this research as an analytic rubric, and holistic criteria, borrowed from the GWT rubric, confirms a serious discrepancy in expectations and opportunity for ESL students facing the GWT.

BEST COPY AVAILABLE

© K. A. Russikoff, 1995

ANALYTIC SCORING PASS/FAIL (P/F) RATIOS

Batch	Distributed Orig. GWT Holistic P/F	ESL Specialists Changes	ESL Specialists P/F Ratio	GWT Raters Changes	GWT Raters P/F Ratio
W	7P:18F	10P:0F	17P:8F	13P:0F	20P: 5F
X	7P:18F	12P:0F	19P:6F	8P:1F	14P:11F
Y	7P:18F	13P:0F	20P:5F	4P:2F	9P:16F
Z	7P:18F	16P:0F	23P:2F	7P:1F	13P:12F
Total:	ESL Specialists:	51 additional Pass; 0 additional Fail		51 altered scores	
	GWT Raters:	32 additional Pass; 4 additional Fail		36 altered scores	
Grand Total:		83 additional Pass; 4 additional Fail		87 altered scores	

One specific example, confirming rater variability based on academic discipline as group membership, was the manner in which rater groups applied the analytic criteria to previously scored GWT essays (Vaughan, 1991). Despite similar training, the ESL Specialists focused more on Language Use, with Content as an important second factor, but GWT Raters focused more on Content. The emphasis on Content by GWT Raters, while supportive of overall faculty response to criteria applied in coursework, is notable in that it was these same raters who expressed surprise at the quality of content and organization in the essays which the analytic scales required these raters to directly assess in addition to other criteria. Their use of the new scales may have presented novel criteria for assessment, contributing to a different and positive emphasis in their scoring.

Rater variability was also confirmed in how the raters applied the scale. Although popular opinion and the higher pass rate for NNS student essays with ESL raters applying the analytic scales (i.e., 51 additional passing papers) may encourage the idea that ESL teachers are more lenient graders, the results of this study do not bear this out. While the mean of factors for ESL readers suggests that ESL Specialists assigned higher scores, this group, as a whole, made more distinctions within their scoring, evidenced by the descriptive statistics which show that they used the full range of possible scores, while GWT Raters were unwilling to assign scores at the lowest end of the range.

DESCRIPTIVE STATISTICS OF ANALYTIC SCORING

FACTORS	ESL Specialists mean	s.d.	range	GWT Raters mean	s.d.	range
Content	16.62	5.187	5-25	14.82	5.143	4-20
Organization	13.61	4.116	4-20	13.3	3.747	4-20
Vocabulary	9.07	2.961	0-15	8.55	2.824	2-15
LanguageUse	9.35	4.067	0-20	7.92	4.151	2-20
Mechanics	11.81	4.437	0-20	10.19	3.848	3-19
TOTAL			24-98			25-100

The use of analytic rubrics compared with holistic rubrics was initially perceived as problematic when differences in pass/fail ratios for analytic scales resulted, since both types of ratings appear to assess the same underlying skills, and both groups of raters appear to be using Language Use and Content as criteria for their assessments.

The most likely explanation of the differences in additional passing papers with the analytic scales is explained by the original batch distributions of 25 previously scored essays in each. Since 19 of the 25 essays in each batch were from the midrange scores of 5, 6, and 7 (emphasizing the 3 score, which ESL students have the most difficulty in overcoming), the higher scores assigned by ESL Specialists may be attributed to these readers' abilities to look beyond typical ESL surface errors (cf. Bochner et al., 1992). And because GWT Raters' results also included 32 additional passing and 4 additional failing essays (from the entire corpus of 100 essays), the findings by Mendelsohn and Cumming (1987) appear applicable to this research also as they maintain that **analytic scales are the best measure for discriminating scores in the midrange.**

© K. A. Russikoff, 1995

Reader Reactions

In addition to the quantitative analyses of holistic and analytic scoring guide usage, an underlying concern for direct writing assessment has always included inconsistency in rater scoring attributed to individual interpretations of the rubrics. In regard to this issue, an unexpected, yet ultimately very important, qualitative concern was presented and must be addressed. Once readers had completed the various stages of the scoring, without solicitation all eight were anxious to share their insights by notes, phone calls, and hallway discussions.

Perhaps the most important awareness was voiced by the senior GWT Rater, one who has been regularly involved in the GWT scoring since its inception on the campus. After completing the analytic scoring, he insisted on discussing his "find," which was that he was surprised to see how strong the content and organization of these ESL papers were because of his new view through the separate analytic factors. In a later separate discussion, a second GWT Rater expressed a similar awareness. This finding confirms that other areas (e.g., Language Use) had previously absorbed these readers' attention to the detriment of true holistic scoring.

With an awareness that holistic assessment may collapse criteria by the one score (and possibly under the weight of one criterion), ESL Specialists had previously expressed unanimous pleasure in rejecting holistic scoring of ESL student writers. They also all agreed that the analytic rubric, while more difficult to manage due to its novelty, was a more appropriate measure of ESL writing. When scoring holistically, they were very concerned that they had been attracted to specific aspects of writing (e.g., "shining examples," "correct use of the word *advice*,"), fearing that their holistic scores probably reflected these superficial elements. With analytic assessment, they were, therefore, pleased to have scores for assignment to these individual factors that would not disproportionately influence the overall score.

One ESL Specialist wrote: "I'm not sure if this is the point of the two rubrics, but the Analytic Rubric seems to be a fairer way to score ESL writers than the GWT Rubric."

© K. A. Russikoff, 1995

Revised version (fewer points but retaining weighted percentages) 4/95

CONTENT:		**1-5**
SCORE	5-4	EXCELLENT TO VERY GOOD knowledgeable • complete thesis development • relevant to assigned topic • may be originally or factually supported
	3-2	GOOD TO AVERAGE some knowledge of subject • limited thesis development • mostly relevant to topic, but lacks detail
	1	FAIR TO POOR limited knowledge of subject • little substance • inadequate development of topic
ORGANIZATION:		**1-4**
SCORE	4	EXCELLENT TO VERY GOOD well-organized • ideas clearly stated/supported • logical sequencing • cohesive
	3-2	GOOD TO AVERAGE loosely organized but main ideas stand out • logical but incomplete sequencing
	1	FAIR TO POOR ideas confused or disconnected • lacks logical sequencing
VOCABULARY:		**1-3**
SCORE	3	EXCELLENT TO VERY GOOD sophisticated range • effective word/idiom choice and usage • appropriate register for author's purpose/audience needs
	2	GOOD TO AVERAGE adequate range • occasional errors in word/idiom form, choice, usage, including unnecessary additions/omissions • but meaning not obscured
	1	FAIR TO POOR limited range • frequent errors of word/idiom form, usage, choice • meaning confused or obscured
LANGUAGE USE:		**1-4**
SCORE	4	EXCELLENT TO VERY GOOD effective complex structures • few errors of agreement, verb tense, articles, pronouns, prepositions
	3-2	GOOD TO AVERAGE effective but simple constructions • several errors of agreement, verb tense, articles, pronouns, prepositions • but meaning seldom obscured
	1	FAIR TO POOR major problems in constructions • frequent errors in agreement, tense, articles, pronouns, fragments, run-ons • meaning confused or obscured
MECHANICS:		**1-4**
SCORE	4	EXCELLENT TO VERY GOOD demonstrates mastery of conventions • few errors of punctuation, spelling, capitalization
	3-2	GOOD TO AVERAGE occasional errors of punctuation, spelling, capitalization • but meaning not obscured
	1	FAIR TO POOR frequent errors of punctuation, spelling, capitalization • poor handwriting • meaning confused or obscured

© K. A. Russikoff, 1995

Brief References

Bereiter, C. (1980). Development in writing. In L.W. Gregg & E.R. Sternberg (Eds.), Cognitive Processes in Writing. Hillsdale, NJ: Erlbaum.

Blok, H. & de Glopper, K. (1992). Large scale writing assessment. In L.Verhoeven & J. H.A.L. DeJong (Eds.), The Construct of Language Proficiency: Applications of Psychological Models to Language Assessment. Philadelphia, PA: John Benjamins Publishing Co.

Bochner, J.H., Albertini, J.A., Samar, V.J., & Metz, D.E. (1992). External and diagnostic validity of the NTID Writing Test: An investigation using direct magnitude estimation and principal components analysis. Research in the Teaching of English, 26 (3),299-314.

Borowiec, E.J. (9188). Pathways to Quality through Diversity: The CSU GWAR in Transition. Long Beach, CA: Office of the Chancellor,California State University.

Chancellor's Office (1987). Executive Order 514, Competency in Student Writing Skills. Long Beach, CA: Office of the Chancellor, California State University.

Chancellor's Office (1982, 1990)Survey of CSU Upper Division Writing Proficiency Requirement. Long Beach, CA: Office of the Chancellor, California State University.

Cherry, R.D., & Meyer, P.R. (1993).Reliability issues in holistic assessment. In M.M. Williamson and B. A. Huot (Eds.), Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations.(pp.109-141). Cresskill, NJ: Hampton Press, Inc.

Collier, V.P. (1994). Table discussion at 1994 TESOL Conference, Baltimore, MD.

Cronbach, L.J. (1988). Five perspectives on the validity argument. In W. Wainer & H.I. Braun (Eds.), Test Validity (pp. 3-127). Hillsdale, NJ: Lawrence Erlbaum.

Cumming, A. (1990). Expertise in evaluating second language compositions. Language Learning, 7 (1), 31-51.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. Review of Educational Research, 49, 222-251.

Davidson, F. (1991). Statistical support for training in ESL composition. In L. Hamp-Lyons (Ed.), Assessing Second Language Writing in Academic Contexts (pp.155-166). Norwood, NJ: Ablex.

Diedrich, P.B. (1974). Measuring Growth in English. Urbana, IL: National Council of Teachers of English.

Diedrich, P.B., French, J.W., & Carlton, S.T. (1961). Factors in judgments of writing ability. (Research Bulletin 61-15). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. 002-172).

Elbow, P. (1993). Ranking, evaluation, and liking: Sorting out three forms of judgment. College English, 55 (2), 187-206.

Fein, D.M. (1980). A comparison of English and ESL compositions. Unpublished master's thesis, University of California, Los Angeles.

Greene, J. A. (1985). Graduation Writing Test Program Review: Final Report. Pomona, CA: California State Polytechnic University.

Hamp-Lyons, L. (1991). Assessing Second Language Writing in Academic Contexts. Norwood, NJ: Ablex.

Henning, G. (1991). Issues in evaluating and maintaining an ESL writing assessment program. In L. Hamp-Lyons(Ed.),Assessing Second Language Writing in Academic Contexts (pp.279-292). Norwood, NJ: Ablex.

Horowitz, D. (1991). ESL writing assessments: Contradictions and resolutions. In L. Hamp-Lyons (Ed.), Assessing Second Language Writing in Academic Contexts (pp.5-18). Norwood, NJ: Ablex.

IRA/NCTE Joint Task Force. (1994). Standards for the Assessment of Reading and Writing. International Reading Association and National Council Of Teachers of English. Urbana, IL: Author.

© K. A. Russikoff, 1995

Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V.F., & Hughey, J.B. (1981). Testing ESL Composition: A Practical Approach. Rowley, MA: Newbury House Publisher, Inc.

Johns, A.M. (1991a). Interpreting an English competency exam: The frustrations of an ESL science student. Written Communication, 8 (3), 379-401.

Leki, I. (1992). Understanding ESL Writers: A Guide for Teachers. Portsmouth, NH: Boynton Cook, Heinemann.

McGirt, J.D. (1984). The effect of morphological and syntactic errors on the holistic scores of native and non-native compositions. Unpublished master's thesis, University of California, Los Angeles.

Mendelsohn, D., & Cumming, A. (1987). Professors' ratings of language use and rhetorical organization in ESL compositions. TESL Canada Journal, 5 (1), 9-26.

Oller, J.W., Jr., (1979). Language Tests at School. London: Longman.

Quellmalz, E. (1980). Problems in Stabilizing the Judgment Process. (CSE Report No. 1 136). Los Angeles: University of California.

Ross, S., Burne, K.G., Callen,J., Eskey, D., McKay, J. (1984). Expectations and evaluations of the second language student: Matters of articulation in California education. A Report to the English Liaison Committee of the Articulation Conference of California.

Vaughan, C. (1991). Holistic assessment: What goes on in the reader's mind? In L. Hamp-Lyons (Ed.), Assessing Second Language Writing in Academic Contexts (pp. 111-126). Norwood, NJ: Ablex.

White, E.M. (1985). Teaching and Assessing Writing. San Francisco: Jossey-Bass Publishers.

Williamson, M.M. (1993). An introduction to holistic scoring: The social, historical and theoretical context for writing assessment. In M.M. Williamson and B. A. Huot (Eds.), Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations (pp. 1-44). Cresskill, NJ: Hampton Press, Inc.

© K. A. Russikoff, 1995